Electrostatic Multipole Representation of a Polypeptide Chain: An Algorithm for Simulation of Polypeptide Properties

ROHIT V. PAPPU, WILLIAM J. SCHNELLER,* and DAVID L. WEAVER†

Molecular Modeling Laboratory, Department of Physics, Tufts University, Medford, Massachusetts 02155

Received Received 21 April 1995; accepted 28 August 1995

ABSTRACT

A method for describing a polypeptide chain based on an electrostatic multipole representation is introduced. The main features of the description are outlined. Appropriate energy functions for nonbonded interactions are developed. The full atomic representation may be retrieved from the electrostatic multipole representation at any point in a calculation. The multipole description and the energy functions are tested by calculation of steric maps for different amino acid side-chain groups. The ability to calculate energetically stable structures is demonstrated by energy conformation maps and the results of energy calculations in optimal secondary structural elements. Results from dynamics simulations of helical chains of polyglycine, polyalanine, polyvaline, and a 21-residue helix obtained from the crystal structure of sperm whale myoglobin are included to demonstrate the efficiency of the algorithm. It is demonstrated that this description of the polypeptide chain is both simple and complete and will allow for the rapid simulation of chain dynamics without loss of essential information about the chain.

Introduction

The purpose of the genetic code is to direct the cellular machinery in living systems to produce required proteins. The production of proteins translates a 1-dimensional alphabet of base-pair triplets into the 3-dimensional order of native-structure proteins in solution, with the structure and internal dynamics determining the biological function of the protein.[1] The information for specifying the 3-dimensional folded structure of a naturally occurring protein is contained in the linear sequence of amino acids produced by the cell, which spontaneously assumes the native structure under appropriate solution conditions.[2] Although

* Present address: Scientific Solutions Inc., 158 Concord Road, Suite E-4, Essex House, Billerica, MA 01821.

† Author to whom all correspondence should be addressed.

CCC 0192-8651 / 96 / 081033-23

the folding process takes place naturally both *in vivo* and *in vitro*, it is difficult to determine the mechanism of folding and the properties of folding intermediates and the native structure. This is partially because of the enormously larger number of naturally occurring proteins ($\sim 10^9$) compared to the data base of determined 3-dimensional structures ($\sim 10^3$),[3] and partially because the intrinsic complexity of multibody problems makes it difficult to solve their dynamics. Nevertheless, it is essential to make progress on the folding prediction problem, both to improve our knowledge and understanding of the natural proteins and to assist in designing and characterizing artificial proteins for which there will be an almost countless variety produceable by modern protein engineering techniques.

In this article we approach the folding problem with the computational strategy suggested by the diffusion–collision (DC) model of protein foldings.[4,5] The DC model suggests a way to circumvent a seemingly intractable problem in the dynamic process of folding a protein from a random coil set of conformations, namely, that there are far too many possible random conformations accessible to the polypeptide chain to find the correct (native) structure on a biologically meaningful time scale. The solution is to reduce the number of bodies relevant to the early folding dynamics to a set of microdomains whose physical properties determine the characteristics of the folding pathways. Since the DC model has recently been reviewed,[5] we will not describe it here. Of particular importance to the present work is the suggestion of the DC model that a three-stage approach to the computational determination of the 3-dimensional structure of a natural or artificial protein be pursued. The first stage is the identification of the microdomains. The second stage is the DC coalescence of the microdomains to a loosely associated molten-globule-like state at the end of the kinetic folding pathway. The third stage is molecular dynamics and energy minimization to attain the interdigitation required for close association of the hydrophobic core. One of the implications of the three-stage approach is that different scales of description of the protein molecule are appropriate for different stages of folding. In particular, a detailed, all-atom description is probably not required for the DC coalescence stage, which would cover most of the folding time scale in many folding proteins. In fact an all-atom description would require very long simulation times for existing computer technology. On the other hand, an all-atom description is appropriate for the interdigitation third stage of folding.

It is the purpose of the present study to describe a representation of a polypeptide chain capable of modeling its static and dynamic aspects using detailed electrostatic nonbonded interactions in a multipole representation. This is accomplished by a rational grouping of atoms in the chain with each group assigned appropriate electrostatic multipole moments. The moments arise from the unequal sharing of electrons between atoms that participate in a covalent bond (due to the different electronegativities of the atoms involved) which results in a partial charge separation associated with the bond. Biologically important atoms, such as O, N, C, S, H, and P possess different electronegativities.[6] Partial charge separation causes the formation of small permanent electric dipoles. A complete set of partial charges for the 20 amino acids was obtained from the work of Burley and Petsko.[6] When we group more than two atoms together as a virtual interaction center, it becomes necessary to include moments higher than dipoles. We describe the polypeptide chain using an electrostatic multipole expansion based on the number of atoms that constitute a group. We adopt an approach related to that of Head–Gordon and Brooks[7] to describe the polypeptide chain, truncating the electric multipole moments at the quadrupole. Our method retains the important torsional degrees of freedom of the backbone. The side chain is modeled as a single interaction center attached to the α-carbon (CA) unit of the backbone. Importantly, we demonstrate that the term "simplified representation" is a misnomer in the context of our work, since we retain the ability to retrieve the full atomic representation at any point in a calculation (with the side-chain atoms at a fixed orientation relative to the backbone atoms) as will be required in the transition from the second to the third stage in the folding algorithm described above. An accurate description in terms of the electrostatic multipole expansion might actually be more effective than attempts at using a full atomic representation to simulate early folding dynamic processes. We suggest an approach that is both simple and complete, in that the details of the chain are reduced for the purpose of rapid simulation, but are retrievable during or after the simulation.

Previous attempts[7] to include the role of polar interactions in a simplified polypeptide rely on the calculation of virtual centers for reducing the number of interaction sites, and rigid bodies for retaining the hard degrees of freedom. The nature of our

description provides for interaction sites, and not a combination of sites and rigid bodies. Where there are more than two atoms in a group, we calculate multipole moments that are higher than dipoles. For dynamical simulations we propose to use an approach quite different from the methods that employ the conventional Fourier decomposed torsional potential.[8] A novel algorithm introduced by Gronbech–Jensen and Doniach is implemented to increase the time step in the dynamics simulations.[9] We demonstrate the correctness of our representation and the important role that nonbonded energy terms play in determining allowed conformations for different side chains, by generating conformational maps for a series of polypeptide sequences. It is believed that hydrogen bonds can be adequately modeled as a sum of dipolar and charge transfer interactions. The electrostatic energy of interaction between the main-chain dipole units in our representation have been calculated and compare favorably with standard hydrogen bonding distances and energy values. In the next section, we detail the procedure used for the multipole parametrization. This is followed by a description of the energy functions that contribute to the potential of mean force due to nonbonded interactions. Next, we present results of the tests employed to validate the multipole representation. We conclude with a discussion of future applications.

Electrostatic Multipole Representation (EMR) of a Polypeptide Chain

Many simulation methods applied to proteins rely on the use of a simplified version of the polypeptide chain. Such simplifications have often been necessary, given the size and complexity of the systems being simulated. Decisions regarding the extent of simplification are usually correlated with the type of information being sought. For example, in simulating ligand binding, the full atomic representation is used to model the active site, while a simplified average representation is used for the rest of the molecule.[1] Simplification usually involves a thermal average over some native degrees of freedom, the remaining degrees of freedom being believed to be most important to the dynamics of the system. Among the more popular approaches to simplification are those developed by Levitt[8] and later modified by McCammon et al.[10] In this method[8, 10] each amino acid residue is treated as a single spherical interaction center in an effective potential that approximates the potential of mean force in solution. The size of the interaction sphere is determined using the van der Waals radii for the different side chains. The 20 naturally occurring amino acid side chains are modeled as 20 spheres of different radii and masses. The intramolecular interactions are chosen to approximate the native interactions in a full atomic representation of the polypeptide. The model allows for the rapid calculation of some dynamical events, such as local rigid body dynamics and certain larger scale motions. The heteropolymeric nature of a polypeptide chain is incorporated via the varying sizes of the interaction spheres. Interactions between nonbonded pairs of interaction centers are given by central force potentials that represent excluded volume and net attractive effects in aqueous surroundings. The scheme adopted for the simplification requires the calculation of an effective dihedral angle in terms of the native torsional angles.[11] This leads to connectivity potentials that depend explicitly on the effective backbone dihedral angle.

The Levitt[8] and McCammon et al.[10] methods involve averaging over some details, and creation of virtual interaction centers that are essentially groups of atoms. The methods save computational complexity and time by not having to distinguish between conformations that differ in finer details. However, it is clear that the extent of these simplifications distorts the nature of the polypeptide backbone. In contrast, we have developed a description that retains most of the important characteristics of the polypeptide chain, such as the native backbone dihedral angles. This ensures reasonable comparison of the results of dynamical simulations with data from spectroscopic and imaging experiments. We believe that it is possible to retain the essential features of the polypeptide, while at the same time having rapid simulation of chain dynamics. We proceed below by first identifying the predominant nonbonded interactions in globular proteins and establishing the role these interactions play in structure stabilization. These nonbonded interactions form the basis of our polypeptide chain description and the development of our energy functions.

Burley and Petsko[6] have reviewed the subject of noncovalent interactions in proteins, particularly the role played by "weakly polar" interactions. They concluded that the 3-dimensional structure of a protein is stabilized by hydrophobic

interactions and by a wide range of electrostatic interactions. The electrostatic interactions result from partial charge formation along a polypeptide chain and involve electric monopole, dipole, and quadrupole moments. In addition, the hydrophobic effect can be explained as being due to the entropically favorable association of nonpolar groups in water. The subsequent enthalpically favorable interactions with the solvent are second-order interactions between the induced moments of clusters of polarizable groups. It has been demonstrated[12] that the hydrogen bonds and salt bridges that stabilize protein secondary and tertiary structure can be accurately modeled as charge–charge, charge–dipole, and dipole–dipole interactions.[6] This model accounts for the directionality of hydrogen bonds. The most significant contribution of hydrogen bonds in proteins is the stabilization of secondary structural elements, for example, the naturally occurring hydrogen bonds of the N–H ··· O type down that backbone of an α helix. It is also known that each peptide unit possesses a significant dipole moment estimated from microwave measurements to be about 12.3×10^{-30} Cm.[12] This dipole moment is comparable to the dipole moment of a water molecule. It is believed that the peptide dipoles add up in an α helix to give a macroscopic dipole moment over the length of the helix.[13] Hol et al.[14] and Sheridan and Allen[15] have calculated an average macroscopic helical dipole to be modeled by a positive and negative charge, each with magnitude of half an electric charge at the N and C terminals of a helix. More recently Zhou et al.[16] demonstrated the role of interhelical ionic interactions in controlling protein folding and stability as studied in a *de novo* designed synthetic two standard α-helical coiled coil. Thus, noncovalent polar interactions play a very important role in protein structure and stability.

In our description, the groupings of the atoms of the polypeptide backbone have been chosen to maintain a balance between reduction of interaction sites and preservation of the important degrees of freedom. Figure 1a shows the full atomic representation of a simple polyalanine chain. The relative orientations of the repeating peptide unit containing the amide nitrogen, the α carbon, the carbonyl carbon, and the attached hydrogens, characterize the polypeptide backbone. We assume that the bond lengths between atoms and the bond angles between three contiguous atoms do not vary much from their equilibrium values. This leaves the backbone dihedral angles as the flexible degrees of freedom. In most proteins, the C′—O

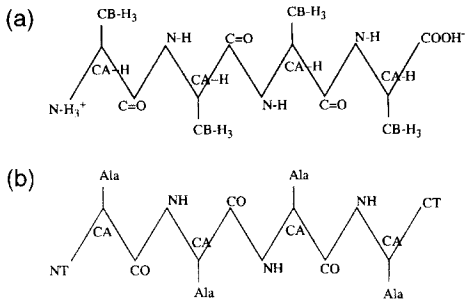


FIGURE 1. The geometry of the (a) full atomic representation for a polyalanine chain compared to that of the (b) electrostatic multipole representation of a polyalanine chain. In the EMR convention CA represents the α-carbon unit, NH the amide group, CO the carbonyl group, NT the N-terminus, CT the C-terminus, and the standard three letter code for the *i*th side chain.

and N—H units are coplanar. This is equivalent to fixing the angle omega (ω) that describes the torsional rotation about the C′—N bond to be $-180°$, the trans conformation of the peptide unit. The two other backbone torsional angles that describe the flexibility of the backbone atoms are the ϕ and ψ angles. They specify rotations about the N—CA and the CA—C′ bonds, respectively. Allowed ranges for the ϕ and ψ angles for different dipeptide conformations were calculated by Ramachandran and Sasisekharan[17] by analyzing various crystal structures and developing empirical rules for limits on the closest contact distances of different atoms. Below, we outline an algorithm for describing the chain that retains the backbone dihedral angles and preserves the native torsional degrees of freedom about the N—CA and CA—C′ bonds, while reducing the number of interaction sites.

The EMR of an n residue polypeptide chain has $4n$ interaction centers according to the following algorithm.

1. Each rigid peptide unit contains two interaction centers for the C′—O and N—H groups. These are treated as colinear dipoles, and the dihedral angle ω is held near $-180°$ to reflect the colinearity. The carbon and oxygen atoms of the carbonyl group possess equal and opposite charges forming a permanent electric dipole. The charges on the nitrogen and hydrogen atoms of the amide group are modified to be equal and opposite, so they too can be treated as a permanent dipole.

Thus, the two components of the peptide unit are modeled as dipole vectors centered at the location of the carbonyl carbon and the amide nitrogen, respectively.

2. For most side chains the CA possesses no charge, whereas the hydrogen atom attached to the α carbon carries a net positive charge of magnitude 0.1 e. Therefore, the CA and associated hydrogen atom are modeled as a single interaction center located at the CA position and possessing a net charge due to the hydrogen atom.
3. The size of the side chains varies depending on the number of atoms that are part of the side chain. For example, the glycine side chain is a single hydrogen atom. The other side chains have different groups attached to the CB (β-carbon) atom. These groups are nonpolar, charged, or uncharged depending on the constituent atoms, their spatial arrangement, and the multipole moments associated with these groups. Some side chains are classified as sulfur containing or aromatic depending on the presence of sulfur atoms or benzene rings. Nonpolar groups have very small permanent multipole moments, while charged groups have significant charges and polar atoms associated with them. In the EMR we treat each side chain attached to the CA as a single interaction center located at the center of mass of the side-chain group. For small side chains, this position is very close to the center of the CB atom; but for bulkier side chains the distance between the CA and the center of mass is different from the CA–CB distance. We describe the side-chain unit using the partial charges and relative positions of each of the atoms in the side chain and calculate multipole moments up to the quadrupole moment. The values of the side-chain χ angles used to generate the multipole moments are shown in Table I. It must be noted that the chosen torsional orientations reflect a choice to keep hydrophobic groups closely packed and charged groups relatively extended. This is done to optimize side-chain packing and inter-side-chain interactions.

The side-chain representation is a gross simplification on the surface. However, closer inspection will reveal that the contributions to the force field are detailed and complete for the following reasons: the contributions of the electronic anisotropy to the side-chain center due to the arrangement of the individual atoms in the side-chain center is not lost in the reduction due to the calculation of

TABLE I.
Torsional Angles about Different Bonds for Side Chain Rotamers Used in EMR Side-Chain Construction.

Side Chain	Bond	Angle	Bond	Angle	Bond	Angle	Bond	Angle	Bond	Angle
Gly										
Ala	CA—CB	57°								
Val	CA—CB	57°	CB—CG1	120°	CB—CG2	−60°				
Leu	CA—CB	57°	CB—CG	−60°	CG—CD1	120°	CG—DC_2	−60°		
Ile	CA—CB	57°	CB—CG1	120°	CB—CG2	−60°	CG_1—CD	−120°		
Ser	CA—CB	57°	CB—OG	−60°						
Thr	CA—CB	57°	CB—CG	−120°	CB—OG	180°				
Cys	CA—CB	57°	CB—SG	−120°						
Asp	CA—CB	57°	CB—CG	−120°						
Glu	CA—CB	57°	CB—CG	−120°	CG—CD	−120°				
Asn	CA—CB	57°	CB—CG	−120°	CG—ND2	−90°				
Gln	CA—CB	57°	CB—CG	−120°	CG—CD	180°	CD—NE_2	−90°		
Arg	CA—CB	57°	CB—CG	−120°	CD—NE	180°	NE—CZ	0°	CZ—NH_2	−90°
His	CA—CB	57°	CB—CG	−120°						
Phe	CA—CB	57°	CB—CG	−120°						
Tyr	CA—CB	57°	CB—CG	−120°						
Trp	CA—CB	57°	CA—CG	−120°						
Met	CA—CB	57°	CB—CG	−120°	CG—SD	180°	SD—CE	0°		

The orientations were chosen to optimize side-chain packing and inter-side-chain interactions.

higher order multipole moments; and calculation of such higher order moments allows the simulation of directional electrostatic effects that contribute to side-chain packing and side-chain orientations. We shall demonstrate the reasonableness of this parametrization by considering side chains from different groups that vary in size and content.

4. The N-terminal (amino group) and the C-terminal (carboxyl group) are treated as separate units centered on their respective centers of mass. In some cases we replaced the terminal carboxyl group with a methyl group which has less pronounced electrostatic properties. This was done to study the interactions of interior side chains in small peptides without the dominating effects of the C-terminal.

Thus, for an n-residue chain we have $2(n-1)$ peptide interaction centers, n CA interaction centers, n side-chain interaction centers, and two distinct terminal interaction centers for a total of $4n$ interaction centers. A schematic representation of the EMR generated for a polyalanine chain is shown in Figure 1b. Each of the centers possesses a calculated multipole moment/moments. The CA center is similar to conventional united atoms.[18] This center is assigned a charge of +0.1 e (see Tables I and II). The two centers that constitute the rigid peptide unit are assigned dipole moments and treated as vectors centered at the locations of the original C′ and N atoms of the backbone. The two dipoles have different magnitudes, the CO unit ($\mu \approx 0.65$ e Å) and the NH unit ($\mu \approx 0.25$ e Å). The EMR associates a dipole moment with an interaction center made of two heavy atoms in a bond with partial charge separation. Since all side chains, except for glycine, possess four or more atoms, and each side chain is described as a single interaction center, higher order moments are calculated for these centers. Here we truncate the calculation of the multipole moments for the side-chain centers at the quadrupole.

CONSTRUCTION OF EMR

The first three backbone centers are placed using standard values for the CA—NH, CA—CO single bond, bond lengths, and the NH—CA—CO bond angle.[12] For each of the remaining interaction centers of the backbone, the coordinates of the previous three centers determine a reference frame. In this frame, the internal coordinates of the center i being added are spherical coordinates relative to center $i-1$, which is the origin. The spherical coordinates are then converted to Cartesian coordinates. The basis vectors of this reference frame are used to transform the Cartesian coordinates from the local frame to a global simulation frame. The vector is rotated and added to the center $i-1$. Side chains are handled in a similar manner, except that the NH, CA, and CO centers nearest to the side chain are used to determine the basis vectors for the local frame.

GENERATING FULL-ATOMIC REPRESENTATION FROM EMR

Internal coordinates giving information about the relative positions of different atoms in the side chains are used to retrieve the full atomic information from the EMR. The amide nitrogen (N), the CA, and the carbonyl carbon (C′) atoms define a local frame of reference. Cartesian coordinates are determined in an empirical lab frame of reference and then transformed into the local reference frame generated by the N, CA, and C′ atoms. Since the positions of these three atoms are fixed with respect to each other, it is possible to generate transformation matrices between this local frame and the lab frame used to specify the full atomic coordinates. If $\mathcal{T}$ defines the transformation between the local and lab frames, then the inverse of $\mathcal{T}$ specifies the transformation from the lab to the local frame.[19] The relative positions of the constituent atoms of a side-chain group are kept fixed, and the orientation of the EMR side-chain interaction center remains fixed with respect to the position of the associated CA unit, i.e., the single side-chain dihedral angle χ_{EMR} is kept fixed during simulations since it is believed that side-chain flexibility does not contribute significantly to rapid dynamical or energy minimization schemes. The orientations of the individual atoms that make up a EMR side-chain group are shown in Table I.

Let $\mathbf{X}_i - \mathbf{C}_i^{\alpha} = \mathbf{R}_{lab}$, be the location of each atom in a lab frame. Here the $\mathbf{X}_i$ are Cartesian coordinates of the atom i, C^{α} is the location of the α carbon, and $\mathbf{R}_{lab}$ is the position vector of the atom i with respect to the α carbon. To calculate the coordinates in a local frame of the EMR defined by the N, CA, and C′ atoms, we perform a transformation of the form, $\mathbf{R}_{local(i)} = \mathcal{T}^{-1}\mathbf{R}_{lab(i)}$ and save

TABLE II.
Sample Multipole Parameters for Polypeptide Main-Chain and Side-Chain EMR Groups in Electronic Units.

EMR Unit	q	μ_x	μ_y	μ_z	Q_{xx}	Q_{xy}	Q_{xz}	Q_{yy}	Q_{yz}	Q_{zz}
NT	+0.40	0.00129	0.3582	−0.1471	0.2390	−0.00032	0.00121	0.2080	−0.09419	0.05399
CA	+0.10									
CO	+0.00	−0.5891	0.03856	−0.33239						
NH	+0.00	0.21639	0.00185	0.103116						
CT	−0.55	0.05756	−0.01017	−0.07684	−0.12495	0.068275	−0.07683	−1.47686	0.044705	−0.04725
Gly	+0.00									
Ala	+0.00	0.030269	0.052383	0.020918	0.087209	−0.03605	−0.01394	0.047535	−0.02457	0.099452
Val	+0.00	−0.414551	0.284281	0.114647	0.671369	−0.248546	−0.022116	0.430567	−0.250375	1.152899
Leu	+0.00	0.11564	0.11003	0.079347	1.506606	−0.07413	−0.18592	0.46779	−0.015425	1.636683
Met	+0.00	−0.36114	0.049885	−0.279350	0.909743	0.437651	0.404516	2.136272	0.268737	0.604844
Ile	+0.00	−0.00666	0.040893	0.165717	0.315958	−0.059695	−0.170076	0.160787	−0.041649	0.513835
Ser	−0.40	0.983683	−0.843719	0.711685	−1.328597	0.784093	−0.459195	−0.633631	0.555599	−1.043271
Cys	−0.05	0.143086	−0.091369	0.103128	−0.211491	0.117733	−0.113939	−0.076783	0.084014	−0.136802
Thr	−0.40	0.170426	0.056144	−0.321158	0.297994	−0.135231	0.089381	0.247510	−0.170010	0.445722
Asn	+0.00	−1.136932	0.460005	0.323658	2.086904	−0.745409	0.089685	0.918252	−0.691554	3.083365
Gln	+0.00	−0.287607	−0.774841	−1.307922	4.829556	1.345327	0.490237	2.453886	0.279598	2.810419
Asp	−1.60	1.567623	−1.120892	1.666808	−3.5652264	1.802635	−1.655943	−1.121101	1.858631	−5.132062
Glu	−1.00	0.638794	0.945352	3.397774	−2.751494	0.308976	−1.890471	−2.967701	−0.901764	−9.841317
Lys	0.65	−0.189481	−0.014696	−0.132268	3.894576	1.587208	1.313328	11.196488	0.921364	2.899898
Arg	+0.55	−0.549130	−1.612110	0.061513	6.247375	4.740264	0.197525	11.905481	−2.003972	6.994100
His	−0.00	0.029325	−0.064427	0.126291	−1.005214	0.376462	−0.281672	−0.176834	0.230406	−0.674788
Phe	−0.10	−0.194020	0.335137	−0.491829	4.847093	−1.580282	1.605778	0.984221	−1.191941	2.380511
Tyr	−0.60	3.714108	−2.88553	5.373414	−23.4288	8.412739	−10.1035	−5.18698	7.054427	−14.0594
Trp	+0.05	−0.638753	0.254195	0.060239	5.354948	−1.457612	−0.420691	0.814622	−1.922687	10.850264

According to the labeling convention used in the EMR algorithm, the N-terminus is denoted by NT, the C-terminus by CT, a methyl group at the C-terminus by MT, the α carbon by CA, and each side-chain unit by the standard three letter form. Peptide amide and carbonyl groups are denoted as NH and CO, respectively.

the resulting coordinates. Since the transformation matrix is real orthogonal, the inverse of this matrix equals its transpose, i.e., $\mathcal{T}^{-1} = \mathcal{T}'$. In the simulation frame of the EMR, the location of the backbone and side-chain units are specified with respect to a common origin. To generate the full atomic coordinates in this simulation frame, the repeating NH, CA, and CO units in the EMR are used to determine local frames of reference. The full atomic coordinates in the local frame $\mathbf{R}_{\text{local}}$ obtained previously are used to calculate the full atomic coordinates in the global EMR frame of interest. This requires one more transformation of the form, $\mathbf{R}_{\text{EMR}(i)} = \mathcal{T}\mathbf{R}_{\text{local}(i)}$, where $\mathbf{R}_{\text{EMR}}$ is the position vector of the atom of interest in the EMR frame, with respect to the location of the corresponding CA atom.

The complete Cartesian coordinates of an atom i are obtained by an addition of the form $\mathbf{X}_{\text{full}(i)} = \mathbf{C}_i^{\alpha} + \mathbf{R}_{\text{EMR}(i)}$. This method of construction depends on the similarity transformations[19] between coordinate frames characterized by identical orthogonal basis vectors. The retrieval of full atomic coordinates is accomplished through a series of operations no more complicated than multiplication of matrices, since the transformation matrices are real orthogonal. It is therefore unnecessary to save and manipulate the internal coordinate information for the constituent atoms of each EMR unit in order to obtain the full atomic information.

CALCULATION OF ELECTROSTATIC MOMENTS FOR EACH EMR CENTER

A monopole q is calculated by determining the net charge associated with a given unit. A dipole is a vector $\boldsymbol{\mu}$ with Cartesian components μ_x, μ_y, and μ_z. The quadrupole moment Q of a given charge distribution is a symmetric 3×3 tensor. To fully specify this quantity, we need to calculate the six independent components of the tensor. The individual multipole moments are given by the following formulas.

$$\text{monopole: } q = \sum_i q_i \tag{1}$$

$$\text{dipole: } \boldsymbol{\mu}_i = \sum_i q_i \mathbf{r}_i \tag{2}$$

and the components of the quadrupole tensor are

$$Q_{xx} = \sum_i q_i x_i^2, \quad Q_{yy} = \sum_i q_i y_i^2,$$

$$Q_{zz} = \sum_i q_i z_i^2 \quad \text{(diagonal elements)}, \tag{3}$$

$$Q_{xz} = \sum_i q_i x_i z_i, \quad Q_{xy} = \sum_i q_i x_i y_i,$$

$$Q_{yz} = \sum_i q_i y_i z_i \quad \text{(off-diagonal elements)}. \tag{4}$$

In general the EMR unit describing a side chain possesses monopole, dipole, and quadrupole moments. Multipole moment parameters are first calculated in the local frame, defined in the previous section, where the positions of the individual atoms are known. The transformation matrix $\mathcal{T}$ is then used to rotate the parameters into the EMR frame. Sample values indicating the magnitudes of various components of the multipole moments for different backbone and side-chain EMR units, calculated using partial charge information at neutral pH, is shown in Table II. From this table it is clear that higher order moments (dipoles and quadrupoles) are determined by the extent of charge separation and the anisotropic distribution of charges in an EMR unit.

Energy Functions to Simulate Nonbonded Interactions in EMR

The most important interactions that stabilize 3-dimensional structures of globular proteins are long-range nonbonded interactions. They explain the formation and aggregation of secondary structures via interactions between groups of atoms that are not adjacent in the amino acid sequence. In the previous section we detailed a method of construction for a polypeptide chain with fewer interaction units than the full atomic representation, each unit characterized by multipole moments. Using multipole moments to characterize interaction sites helps preserve significant details of the nonbonded interactions. Below we develop a molecular force field that describes the nonbonded interactions between EMR units as being a sum of electrostatic multipole and van der Waals terms.

ELECTROSTATIC MULTIPOLE INTERACTION TERMS

If we place a localized charge distribution $\rho(r)$ in an external electric field due to another charge distribution, the electrostatic interaction potential energy of the system is

$$V = \int \rho(r)\Phi(r)\, dr \tag{5}$$

where $\Phi(r)$ is the potential producing the electric field. If the potential Φ varies slowly over the region where the charge distribution is nonnegligible, the expression for Φ can be expanded in a Taylor series of the form,[20]

$$\Phi(r) = \Phi(0) + \mathbf{r} \cdot \nabla\Phi(0) + (1/2)\left(\sum_i \sum_j r_i r_j \frac{\partial^2}{\partial r_i r_j} \Phi(0) \right) + \cdots \quad (6)$$

where $\Phi(0)$ is the value of $\Phi(r)$ at $r = 0$. This expression can be rewritten as

$$\Phi(r) = \Phi(0) + \mathbf{r} \cdot \mathbf{E}(0) + (1/2)\left(\sum_i \sum_j r_i r_j \frac{\partial}{\partial r_i} E_j \right) + \cdots . \quad (7)$$

Using the definition of the external potential in eq. (7), the interaction energy V in eq. (5) can be written in the form

$$V = q\Phi(0) - \boldsymbol{\mu} \cdot \mathbf{E}(0) - (1/6) \sum_i \sum_j Q_{ij} \frac{\partial}{\partial r_i} E_j(0) - \cdots \quad (8)$$

where q is the spherically symmetric charge or monopole, $\boldsymbol{\mu}$ is the electric dipole moment, and the Q_{ij} are elements of the electric quadrupole moment tensor. The expansion in eq. (8) shows the interaction of the different multipole moments with an external field, i.e., the charge with the potential, the dipole with the field, and the quadrupole with the field gradient.

The total electrostatic energy in eq. (5) is rewritten as a sum of pairwise interactions between the permanent multipole moments of two groups in the form,

$$V = \sum_i \sum_j V_{ij}. \quad (9)$$

In the present case the indices i and j run from 0 to 2 since we truncate the multipole expansion at the quadrupole term. The lowest order terms of this series represent charge–charge interactions and the highest order terms are of the quadrupole–quadrupole type. The nine terms of interest are

$$V_{00} = q_1 q_2 r^{-1} \quad \text{monopole–monopole} \quad (10)$$

$$V_{01} = q(\boldsymbol{\mu}_1 \cdot \mathbf{r}) r^{-3} \quad \text{monopole–dipole} \quad (11)$$

$$V_{10} = q(\boldsymbol{\mu}_2 \cdot \mathbf{r}) r^{-3} \quad \text{dipole–monopole} \quad (12)$$

$$V_{11} = (\boldsymbol{\mu}_1 \cdot \boldsymbol{\mu}_2) r^{-3} - 3(\boldsymbol{\mu}_1 \cdot \mathbf{r})(\boldsymbol{\mu}_2 \cdot \mathbf{r}) r^{-5} \quad \text{dipole–dipole} \quad (13)$$

$$V_{02} = q_1(\mathbf{r} \cdot Q_2 \cdot \mathbf{r}) r^{-5} \quad \text{monopole–quadrupole} \quad (14)$$

$$V_{20} = q_2(\mathbf{r} \cdot Q_1 \cdot \mathbf{r}) r^{-5} \quad \text{quadrupole–monopole} \quad (15)$$

$$V_{12} = 2(\boldsymbol{\mu}_1 \cdot Q_2 \cdot \mathbf{r}) r^{-5} - 5(\boldsymbol{\mu}_1 \cdot \mathbf{r})(\mathbf{r} \cdot Q_2 \cdot \mathbf{r}) r^{-7} \quad \text{dipole–quadrupole} \quad (16)$$

$$V_{21} = (-2)(\boldsymbol{\mu}_2 \cdot Q_1 \cdot \mathbf{r}) r^{-5} + 5(\boldsymbol{\mu}_2 \cdot \mathbf{r})(\mathbf{r} \cdot Q_1 \cdot \mathbf{r}) r^{-7} \quad \text{quadrupole–dipole} \quad (17)$$

$$V_{22} = \tfrac{1}{3} \text{ trace } Q_1 Q_2 r^{-5} - \tfrac{20}{3}(\mathbf{r} \cdot Q_1 Q_2 \cdot \mathbf{r}) r^{-7} + \tfrac{35}{2}(\mathbf{r} \cdot Q_1 \cdot \mathbf{r})(\mathbf{r} \cdot Q_2 \cdot \mathbf{r}) r^{-9}. \quad \text{quadrupole–quadrupole} \quad (18)$$

If all the charges in a system are described explicitly, fully accounting for fluctuations and charge transfer reactions in the presence of a solvent, then a dielectric constant of one will adequately describe the electrostatic interactions. However, since we do not use a full atomic point charge description, distance dependent dielectric functions are used to model screening effects of solvents. Distance dependent dielectric functions ensure that screening effects are minimal at short distances and tend to a constant value at larger distances reflecting the bulk properties of the solvent.

SOLVENT MODELS

An ideal representation for the solvent would be to include water molecules explicitly. Molecular dynamics simulations of short peptides have been performed using variations of the ST2 or TIPS representations for water molecules.[21,22] Any improvement upon the computations involved with such models for the solvent would be small due to the large number of solvent molecules that need to be included for moderate dilution. Other methods of including solvent effects depend upon modeling the screening effects of a solvent using continuous distance-dependent dielectric functions.

Linear forms such as $\varepsilon = 4r$ or $\varepsilon = 8r$, where ε is the dielectric function, give low enough values at short distances, but at larger distances they exceed the bulk value for the dielectric constant of the solvent. Daggett et al.[23] demonstrated that a

sigmoidal distance-dependent form, which is a modification of the function proposed by Hingerty et al.,[24] adequately models the distribution of solvent molecules around a globular protein. The form of this sigmoidal function is

$$\varepsilon(r) = D - \frac{D-1}{2}((rs)^2 + 2rs + 2)\exp(-rs), \tag{19}$$

where D is the plateau value of ε reached at long distances, s is the slope of the sigmoidal segment of the function, and the distance r is in angstroms. For aqueous solutions D is assumed to be 78, and depending on the ionic concentration, the parameter s is chosen to be either 0.3 or 0.16.[23] This and other possible forms for the distance dependent dielectric function are shown in Figure 2. As is seen in this figure, the behavior of the function $\varepsilon = 4r$ is quite similar to that of the sigmoidal function for distances up to 20 Å. Beyond this distance the function $\varepsilon = 4r$ gives values that exceed the value of the bulk dielectric D. We use the sigmoidal function because it accurately underestimates screening effects at short distances and tends to the value for the bulk dielectric constant D at large distances. In addition, unlike simple linear functions such as $\varepsilon = 4r$, the screening properties of the sigmoidal function depends on the nature of the solvent, as determined by the parameters D and s that describe the dielectric properties of the bulk solvent and concentration of ions in the solvent.

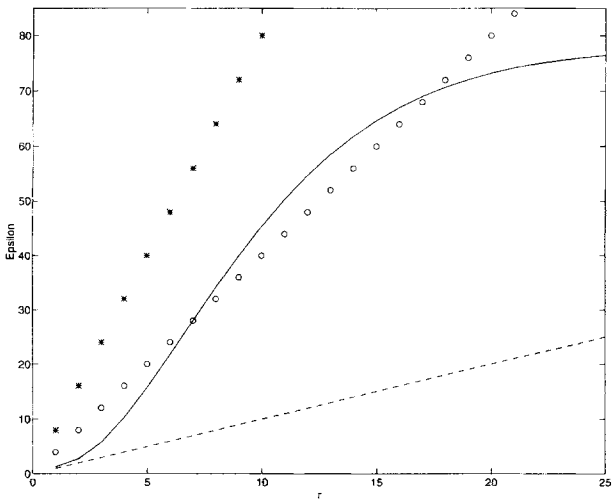


FIGURE 2. Functional forms for different linear and sigmoidal distance dependent dielectric functions. (---) $\varepsilon = r$; (—) sigmoidal function used in this work $\varepsilon(r) = D - (D-1)/2((rs)^2 + 2rs + 2)\exp(-rs)$; (○) $\varepsilon = 4r$; and (*) $\varepsilon = 8r$.

We assume that the effect of a dielectric is similar for all pairwise multipole interactions. This is accurate, since each higher order multipole moment arises due to a specific spatial distribution of point charges. In actual calculations dielectric effects are included only when the center to center distance between EMR units exceeds the sum of their van der Waals (vw) radii and the diameter of a water molecule. This is necessary to exclude the screening of electrostatic interactions when the distance of separation between two EMR units is not large enough to allow a solvent molecule between them. The electrostatic (el) contribution to the molecular force field takes the form

$$V_{\mathrm{el}(ij)} = \frac{1}{\varepsilon(r_{ij})}(V_{00} + V_{01} + V_{10} + V_{11} + V_{02} + V_{20} + V_{12} + V_{21} + V_{22}) \tag{20}$$

and

$$\varepsilon(r_{ij}) = \begin{cases} 1 (r_{ij} < r_{\mathrm{vw}(i)} + r_{\mathrm{vw}(j)} + 2.8\,A) \\ \varepsilon(r_{ij} \geq r_{\mathrm{vw}(i)} + r_{\mathrm{vw}(j)} + 2.8\,A). \end{cases} \tag{21}$$

EXCLUDED VOLUME van der WAALS FUNCTION

Typical 6–12 type van der Waals functions describe a combination of attractive and repulsive interactions. Attractive interactions involve induced electric multipoles, and repulsive interactions prevent unfavorable spatial overlap of interacting units. A variety of exponents from the inverse 12th to the inverse ninth or eighth are used to model the repulsive term.[8] Softer terms corresponding to exponents smaller than 12 have been used for the repulsive term[8] for computational simplicity and overall numerical stability in a calculation. All of the electrostatic interactions, eqs. (10)–(17) (except for quadrupole–quadrupole interactions), between EMR units have a longer range than either of the two terms in a van der Waals interaction. Thus, van der Waals interactions contribute strictly to short-range excluded volume effects.

The van der Waals excluded volume function in its general form is written as

$$V_{\mathrm{vw}(ij)} = \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \tag{22}$$

where r_{ij} represents the van der Waals distance between the two EMR units i and j, A_{ij} and B_{ij} are the repulsive and attractive coefficients, respectively. This is the conventional Lennard–Jones (6–12) type function used to model dispersive interactions. Repulsive and attractive parameters are calculated for all pairs of interacting units. There are a total of 24 different units in our representation for a polypeptide chain. These include the two peptide units, a CA unit, 19 side-chain units (proline is not included in this list), and two distinct terminal groups, leading to a 24×24 matrix of coefficients A_{ij} and B_{ij} for all possible pairwise interactions.

The strength of the attractive interactions between induced dipole moments depends on the static polarizabilities of the interacting units. Larger units composed of distinctly nonpolar atoms, have higher polarizabilities and therefore stronger London attractive interactions.

The attractive coefficients B_{ij} in eq. (20) for dispersive van der Waals interactions can be calculated using the Slater–Kirkwood formula,[25] provided we have accurate values for the static polarizabilities of the interacting groups. The formula is

$$B_{ij} = \frac{3}{2}\frac{eh}{\sqrt{m_e}}\frac{\alpha_i\alpha_j}{\sqrt{\frac{\alpha_i}{N_i}} + \sqrt{\frac{\alpha_j}{N_j}}} \qquad (23)$$

where e is the electric charge, m_e the mass of an electron, and h is Planck's constant. $\alpha_{i(j)}$ are the static polarizabilities of the groups $i(j)$ and $N_{i(j)}$ are the effective values of the number of outer shell electrons, as calculated by Pitzer[26] and Scott and Scheraga.[27]

A quick estimate of the polarizability of any molecule can be obtained by adding the contributions of constituent bonds to the polarizability.[28] This is plausible if we assume that electrons in a molecule can be divided into a number of independent groups such that the total wave function can be factored into a product of a set of functions, each corresponding to a single group of electrons with no correlation between the motions of electrons in any two distinct groups. The additivity of the polarizability is therefore a measure of the separability of a molecule into electron groups, such as inner and outer shell electrons in chemical bonds.[28] The polarizability of a bond is different parallel to the direction of a bond compared to a direction perpendicular to the bond. If an electric field makes an angle θ with a chemical bond, the static polarizability α_θ is given by

$$\alpha_\theta = \alpha_\parallel \cos^2\theta + \alpha_\perp \sin^2\theta. \qquad (24)$$

Averaging over all orientations θ gives the average value of the static bond polarizability, as $\alpha = 1/3(\alpha_\parallel + 2\alpha_\perp)$. The average polarizability of an EMR unit is a sum of the average polarizabilities of its constituent bonds.[28] Wherever possible the different side-chain groups of amino acids in a polypeptide chain are identified as unsaturated analogs of small molecules, i.e., serine is methanol sans a hydrogen, cysteine is methane-thiol sans a hydrogen, and phenylalanine is toluene sans a hydrogen. The average static polarizabilities α are known[29] for these molecules; and since the side-chain units differ from these molecules only in the absence of hydrogen atoms, we postulate that the static polarizabilities for these groups adequately models the static polarizabilities of the side-chain groups. Static polarizability values used for each of the EMR units is shown in Table III.

The values of N in eq. (23) for the van der Waals parameter B_{ij} correspond to the number of outer shell electrons for two atoms involved in a van der Waals interaction. Since each EMR unit is a group of atoms, we formulate an empirical scheme to calculate a value for N that can be used in eq. (23). Scott and Scheraga[27] provided a plot for the variation of an effective number of outer shell electrons N_{eff} with atomic number Z for inert gases (see Fig. 3). We use this plot y to estimate a value for Z (defined here as an effective "atomic number") for each EMR unit and then determine the corresponding value of N_{eff} from the plot.

To calculate the effective atomic number, we postulate that only outer shell electrons participate in the formation of a molecule or EMR unit. Hence the "effective atomic number" is calculated to be the sum of the outer shell electrons due to each of the atoms in the EMR unit. We use this algorithm for amino acid side-chain groups where more than two heavy atoms are part of the group. The algorithm is based on the assumptions that:

1. outer shell electrons for each atom play a dominant role in the formation of a molecule, and
2. the wave function of a molecule can be written as a product of a set of functions, each set describing a group of electrons whose motions are uncorrelated. Values of N_{eff} calculated according to this algorithm using data from Figure 3 are given in Table III.

TABLE III.
Parameters for Calculating van der Waals Coefficients, Using Formulas for A_{ij} and B_{ij} in Eqs. (23) and (25), Respectively.

EMR Unit	Static Polarizability α (10^{-24} cm^3)	Effective No. "Outer Shell Electrons," N_{eff}	van der Waals radii (Å)	Diffusion Coefficient (10^{10} Å^2/s)
NT	2.26	7.94	1.97	14.31
CA	0.79	4.50	1.75	10.66
CO	1.99	9.2	2.15	10.66
NH	0.58	6.13	1.85	12.44
CT	2.72	15.63	2.45	9.61
Gly	0.42	0.90	1.20	24.04
Ala	1.97	6.90	1.97	14.31
Val	5.7	17.0	2.86	8.72
Leu	8.2	19.0	3.25	13.00
Ile	8.0	19.0	2.67	6.00
Cys	3.37	12.37	2.95	8.62
Ser	2.44	12.37	2.65	9.60
Thr	4.35	17.5	3.38	10.42
Met	8.0	19.30	3.55	7.91
Asp	5.1	18.7	2.98	6.00
Glu	6.9	19.80	3.21	7.77
Asn	5.7	18.7	3.19	7.82
Gln	7.8	19.80	3.21	7.79
Lys	13.5	20.08	3.18	7.86
Arg	8.5	25.0	3.48	7.77
His	9.3	20.08	3.18	7.84
Phe	11.82	22.04	3.47	5.73
Tyr	13.1	23.5	3.57	7.00
Trp	18.2	29.5	4.98	10.67

Values for N_{eff} were obtained from the plot in Figure 3. Values of α were obtained from information regarding the static polarizabilities of saturated organic molecules that are analogs of the EMR units. Column four gives the diffusion coefficients for the EMR groups used in the Langevin dynamics simulations [see eq. (28)].

The repulsive parameter A_{ij} in eq. (22) is obtained from the condition that $V_{vw(ij)}$ be a minimum at a distance $r_{ij(min)}$, assumed to be the sum of the van der Waals radii of the interacting groups.[29] A is then given in terms of B for the 6–12 type potential as,

$$A_{ij} = \tfrac{1}{2} B_{ij} r^6 \min(ij). \qquad (25)$$

To determine the A_{ij}, we need to determine appropriate values for the van der Waals radii of the different main-chain and side-chain units. Values for the effective radii have been shown to depend on factors such as the strength of the attractive forces holding the molecule together and the relative orientation of the van der Waals contact.[30] It has also been shown from crystallographic data that van der Waals radii are usually 0.75–0.83 Å greater than single bond covalent radii.[30] This implies that identical atoms involved in a single covalent bond can approach one another much more closely than the sum of their individual van der Waals radii. van der Waals radii for atoms occurring in biomolecules have been published.[27,30] The van der Waals radius of an EMR unit with no more than two heavy atoms is taken to be a sum of the individual single bond covalent radii plus 0.8 Å.[30] For units with more than two heavy atoms, such as side chains larger than glycine and alanine, known values of b,[29] the volume parameter in the van der Waals equation of state, are used to calculate the van der Waals radii. This is in accordance with our previous identification of larger side-chain units as unsaturated analogs of molecules. Since the volume parameter b describes the volume occupied by a spherical van der Waals shell of a molecule of radius r_{vw}, written as

$$b = \tfrac{4}{3} \pi r_{vw}^3. \qquad (26)$$

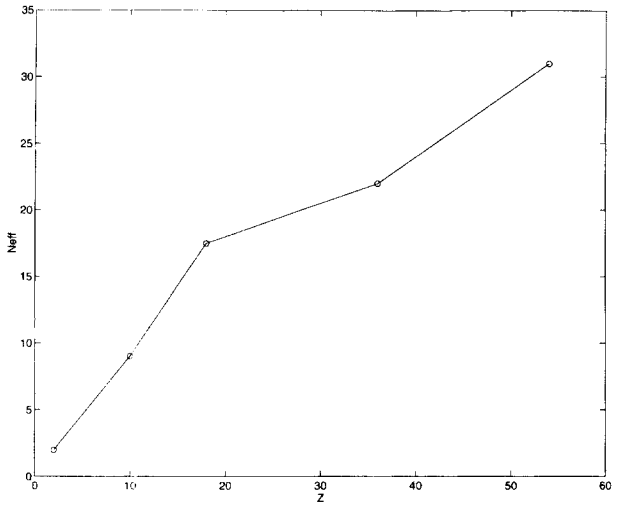


FIGURE 3. Variation of screening constant N with atomic number Z for inert gas atoms, based on work by Scott and Scheraga[27] using data from Pitzer.[26] To calculate N from this plot, an effective "atomic number" Z is determined for each EMR unit based on a simple addition of the number of outer shell electrons due to each of the constituent atoms of an EMR unit.

Thus, the van der Waals radius of an EMR unit in terms of the parameter b is calculated from eq. (26) to be

$$r_{vw} = 3\sqrt{\frac{3b}{4\pi}}. \quad (27)$$

For example, the value of b for the tyrosine EMR unit (similar to that of methyl-phenol[29]) is 190 Å^3, and the corresponding value for r_{vw} obtained from eq. (27) is 3.57 Å.

Bonded Constraints and Langevin Dynamics

In the previous section we developed a molecular force field to calculate the nonbonded (or noncovalent) interactions between EMR units. Bonded constraints are a consequence of covalent interactions between atoms that come together to form the polypeptide chain. Such constraints have been traditionally modeled using harmonic potential functions with stiff springs (to prevent significant deviations from equilibrium values).[18,31,32] In simulations where the time evolution of a dynamical system is modeled using Newton's equations of motion, the Verlet SHAKE algorithm can be used to obtain the constraint forces mathematically.[33] The principle is based on using the *a priori* knowledge of nonbonded interactions to solve for the required constraints by Lagrange's method of undetermined multipliers (which are scalar coefficients).[19]

Recently, Gronbech–Jensen and Doniach developed a novel method of mathematically calculating bonded constraints for dynamical systems modeled using the Langevin equation in the overdamped limit.[9] The procedure involves obtaining the constraint forces by solving a system of linear equations. We have implemented the Gronbech–Jensen and Doniach method to obtain constraint forces required to maintain the polypeptide geometry in the EMR during dynamics simulations. This method of solving for constraint forces mathematically (i.e., calculating Lagrange's multipliers) can be used for any simulation of choice, such as Monte Carlo methods or energy minimization routines. The concept is based on recognizing the interaction centers (or atoms as the case may be) that need to be constrained. Such techniques preclude the need to use steep "potential well" functions that depend explicitly on bond lengths and bond angles.

Motion of a residue or EMR unit is largely determined by the variation of its nonbonded interactions with other units. A discrete Langevin equation that describes the trajectory $\mathbf{r}_i$ of an EMR unit i in the diffusive limit is[9,31,34]

$$\frac{d}{dt}\mathbf{r}_i = \frac{-D_i}{k_B T}\frac{\partial V}{\partial \mathbf{r}_i} + \frac{D_i}{k_B T}\mathbf{R}_i + \frac{D_i}{k_B T}\mathbf{S}_i, \quad (28)$$

where

$$V = V_{el} + V_{vw} \quad (29)$$

V_{el} and V_{vw} are given by eqs. (21) and (23), respectively, and D_i is the diffusion coefficient of the ith unit (see Table III). Equation (28) specifies the time evolution of the EMR unit trajectories, where $\partial V/\partial \mathbf{r}_i$ is the interaction term. $\mathbf{R}_i$ represents the random force exerted on the ith EMR unit by the surrounding solvent molecules. $\mathbf{S}_i$ is the ith component of the constraint force, written as $\Sigma_j s_{ij}(\mathbf{r}_j - \mathbf{r}_i)$, where the coefficients s_{ij} are scalars, defined for two units i and j that are constrained. The constraint term projects out all the components of $\partial V/\partial \mathbf{r}_i$ that alters the constrained distances in the chain.[9] The Langevin equations are integrated using a modification of the algorithm of Ermak and McCammon.[34] We used a fourth-order Rünge–Kutta method to integrate the Langevin

equation[35] instead of the Euler method of Ermak and McCammon.[34] The slight increase in the number of calculations does not increase the computation time significantly, but offers significant numerical stability.

In addition to bond length and bond angle constraints, the peptide unit has to be kept flat, i.e., $\omega = -180°$. This is a purely quantum mechanical artifact that has no classical analog. It is important to maintain this rigidity throughout a dynamical or any other simulation if we are to adequately model the properties of a polypeptide chain. Instead of using mathematical constraints, we use a potential function that depends on the angle ω, as shown in eq. (30) below,

$$V_\omega = K_\omega(\cos \omega + 1) \tag{30}$$

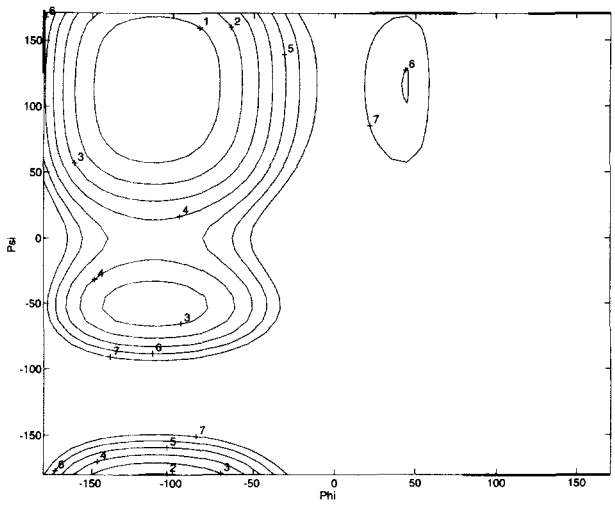
where K_ω is a well-depth parameter. The value for K_ω is directly related to the energy associated with the partial double bond of the peptide unit, estimated from microwave measurements to be approximately 21 kcal/mol.[12]

Higher order multipole moment interactions between backbone elements depend on the orientation of the moments with respect to each other. Since the peptide unit is modeled as two colinear dipoles, the interaction between the NH and CO units is determined by the relative orientation of the dipole vectors. The monopole–dipole and dipole–dipole interactions along the backbone specify the relative orientations of the backbone units as described by the ϕ and ψ dihedral angles. It is important to note that modeling such higher order multipole interactions along the chain precludes the need for using the Levitt torsional energy function that depends explicitly on the backbone ϕ and ψ dihedral angles.[8]

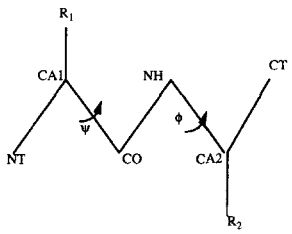
Results

We have developed an algorithm that describes a polypeptide chain in terms of electrostatic multipole moments. Side chains and backbone units are included in this representation. The native backbone degrees of freedom are retained explicitly. Two principal features of the EMR algorithm are the location of EMR units representing groups of atoms along a polypeptide chain and the energy functions that describe the interactions between the EMR units. Amino acid side chains are distinguished in the EMR by differences in size, van der Waals parameters, and multipole moments.

To test the accuracy of the EMR description of polypeptide units, the average steric behavior of EMR units in a polypeptide has to be investigated. The simplest conformations to study are amino acid residues in dipeptides. The conformational energy surface for a residue in a dipeptide is a good representative of the side-chain residue in a polypeptide chain. Ramachandran and Sasisekharan[17] and Scheraga[36] have documented the propensity of different amino acids for adopting sterically favorable conformations, using dipeptides as model units. We have generated steric energy surfaces and 2-dimensional energy contour maps for EMR dipeptides and compared them to previous work.[36,37] Figure 4 is a contour plot of the steric energy for an alanine dipeptide with a carboxyl group at the C-terminus (Fig. 4b). The contours correspond to alanine in well-defined secondary structures. The absolute minimum on the energy surface of Figure 4 is located at $\phi = -107°$ and $\psi = 116°$, which is in the β-sheet region of the surface. This value is in fairly good agreement with the work of Zimmerman et al.[37] who calculated the conformational energy contour map for *N*-acetyl-*N'*-methyl amide, using empirical energy functions. The location of the absolute minimum in their work was at $\phi = -84°$ and $\psi = 72°$. The empirical energy functions they used included specific short-range terms to model hydrogen bonds, and is significantly different from the EMR molecular force field, eqs. (21)–(23). The excluded regions on the steric contour plot of Figure 4 represent barriers that prevent alanine from adopting unfavorable orientations in a polypeptide chain. These steric energy barriers are a consequence of geometrical constraints on the possible values of ϕ and ψ, preventing unfavorable spatial overlap between NH and CO peptide units, and between peptide units and the terminal groups. Specific sterically unfavorable orientations in a dipeptide are described in Table IV. The ridge through $\phi = 0°$ and $\psi = 0°$ is due to the proximity of the NH group to the N-terminus, and CO group to the C-terminus. The ridges at $\phi = 120°$ and $\psi = -120°$ are due to the interaction of the side-chain units with the NH and CO peptide units, respectively. The separation between each contour on this plot is 1 kcal/mol, and the contour lines are labeled with the steric energy in kilocalories per mole above the minimum at $\phi = -107°$ and $\psi = 116°$. The location of the α-helical region, indicated on this map, is 6 kcal/mol above the minimum in the β region. According to this map, only regions of the ϕ–ψ conformational space corre-



(a)



(b)

FIGURE 4. (a) Ramachandran steric energy contour plot for an alanine dipeptide indicating the regular secondary structure regions. The separation between each contour is 1 kcal / mol, and the contour lines are labeled with the steric energy in kcal / mol above the minimum at $\phi = -107°$ and $\psi = 116°$. (b) The geometry of an EMR dipeptide used to generate the steric contour maps in Figures 4–6 showing the ϕ–ψ angles of the contour maps. The choice of the dipeptide is similar to that used by Ramachandran and Sasisekharan.[17] In the EMR convention CA represents the α-carbon unit, NH the amide group, CO the carbonyl group, NT the N-terminus, CT the C-terminus. R_i $(i = 1, 2)$ is any side-chain group for the conformal search.

sponding to structural motifs such as β sheets and α helices fall near local minima in the potential surface. This is in contrast to Figure 5, which is a Ramachandran plot for glycine, for which the steric restrictions on the allowed conformations are not as severe. Since the glycine side chain does not possess an asymmetric center, the Ramachandran plot in Figure 6 is symmetric about the barrier through the $\phi = 0°$ and $\psi = 0°$ line. The results of Figures 4 and 5 show that the hard sphere potential parameters A_{ij} for alanine and glycine keep these side chains from adopting sterically unfavorable conformations in the EMR, and allow the well-known secondary structure motifs seen in all-atom calculation of energy maps. Figures 6a–c is a Ramachandran plot for the remaining set of hydrophobic residues, valine, leucine, and isoleucine. These amino acids differ from alanine in the number of additional CH, CH_2 and CH_3 groups attached to the CB atom. The side-chain interaction center is located at the center of mass of the residue. The CA–side-chain center distance is therefore larger than the original CA–CB distance for each of these groups. The individual atoms in

TABLE IV.
Sterically Unfavorable Orientations and Interactions in EMR Dipeptides.

ϕ	ψ	Description of Dominant Unfavorable Interactions
120°	Any value	Interaction of the second CA unit with the CO group
0°	Any value	Interaction of the C-terminus with the CO group
−120°	Any value	Interaction of the second side chain with the CO group
Any value	120°	Interaction of the first CA unit with the NH group
Any value	0°	Interaction of the N-terminus with the NH group
Any value	−120°	Interaction of the first side chain with the CO group

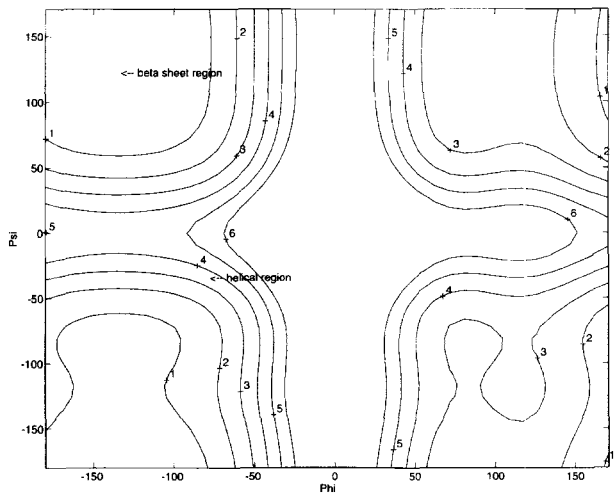


FIGURE 5. Ramachandran steric energy contour plot for a glycine dipeptide. The size of the glycine side chain is reflected in the increase in the percentage of sterically allowed regions. The glycine map is symmetric about the $\phi = 0°$ and $\psi = 0°$ lines.

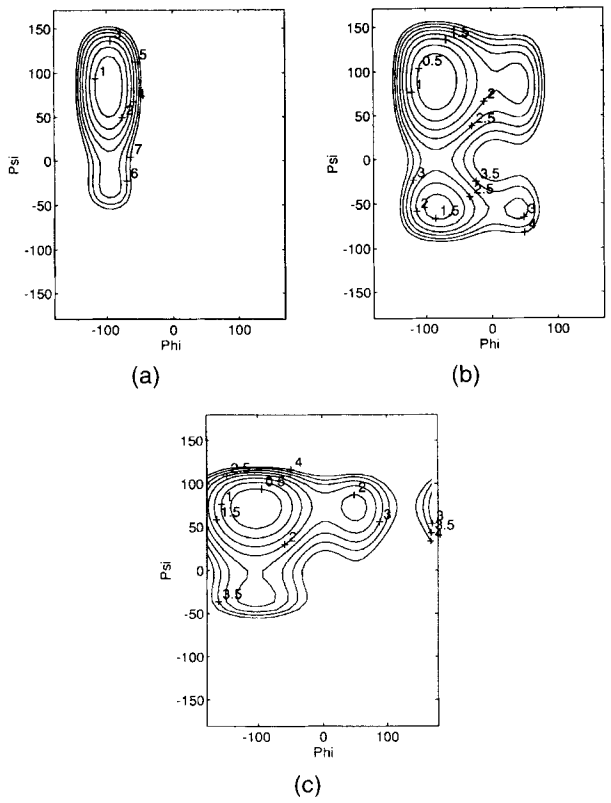


FIGURE 6. (a) Ramachandran steric energy contour plot for a valine EMR dipeptide. (b) Ramachandran steric energy contour plot for a leucine EMR dipeptide. (c) Ramachandran steric energy contour plot for a isoleucine EMR dipeptide. The sterically allowed regions are confined to $\phi-\psi$ angles that define well-defined secondary structural motifs, and are more alanine-like, in that fewer regions of the $\phi-\psi$ space are sterically allowed.

these side-chain rotamers are positioned to obtain close packed conformations. The result is the alanine-like contour maps shown in Figures 6a–c. These steric maps were generated using the steric van der Waals energy function alone to test for the accuracy of the steric parameters A_{ij} and B_{ij} used in the Lennard-Jones function of eq. (23). The alanine-like behavior (i.e., that only well-defined secondary structural motifs seem to be sterically favorable) of the bulkier hydrophobic amino acids indicate that the steric parameters used to model the EMR side-chain interaction sites are reasonably accurate. In simulations implementing the EMR algorithm, the electrostatic term of eq. (21) is included to model the complete electronic anisotropy of the side-chain groups.

Ramachandran plots generated using the EMR parametrization show that the average steric behavior and the detailed electrostatic behavior of dipeptide units are preserved. From the results of Figures 4–6 we conclude that the steric parametrization of the EMR units is sufficiently accurate to reflect the average sterically allowed conformations of amino acids in a polypeptide chain. This makes the EMR useful for the study of the properties of protein segments that are difficult with full atomic representations.

CONTRIBUTIONS OF ELECTROSTATIC ENERGY TO STABILITY OF ELEMENTS OF SECONDARY STRUCTURE

The right-handed α helix is one of the most prominent polypeptide regular structures. It has 3.6 residues per turn and a translation per residue of 1.5 Å.[38] The ideal backbone dihedral angles for the right-handed α helix are $\phi = -57°$ and $\psi = -47°$,[12] corresponding to optimal alignment of the stabilizing hydrogen bonds. The atoms of the backbone pack closely making van der Waals interactions favorable. The backbone carbonyl oxygen of each residue is hydrogen bonded to the backbone hydrogen of the amide group of the fourth residue. These hydrogen bonds are 2.86 Å long and are very nearly parallel to the helical axis.[12]

It is believed that hydrogen bonds in an α helix can be modeled as a sum of dipole–dipole and charge transfer interactions.[6] This formulation is reasonable given that hydrogen bonds are directional. We tested the formulation by calculating the energies and distances for carbonyl carbon and amide EMR units that could be involved in hydrogen bonds in an ideal helix. Alter et al.[39] performed calorimetric measurements using polyvaline at 25°C. Their method is based on estimating the contribution of hydrogen bonds to the change in free energy in the calorimetric measurement. Their results indicate that the energy associated with a hydrogen bond is −1.4 kcal/mol. However, their measurement does not involve an explicit calculation of the pairwise hydrogen bonding energy, but is instead an estimate of the free energy change due to the making or breaking of hydrogen bonds. More recent work has focused on the calorimetric determination of the enthalpy change for the α helix to coil transitions of an alanine peptide in water. The contribution of making and breaking of the hydrogen bond to the enthalpy was found to be 1.3 kcal/mol. In previous work on the dynamics of helical segments, McCammon et al.[10] and Schneller and Weaver[32] used values of −1.4 and −1.3 kcal/mol for the

helix stabilizing hydrogen bonding well-depth parameter, as a contribution to the potential energy for a residue at the helix–coil interface in the single center per residue model. Since it is not possible to model directional interactions in a spherical interaction center model for a polypeptide chain, empirical estimates were used to specify a hydrogen bonding well-depth parameter that confines the virtual dihedral angle to the helical range. The values[10,32] were chosen to conform with the calorimetric measurements. None of the above estimates or measurements explicitly calculates the energy of interaction between atoms in a hydrogen bond pair.

We studied a 10-residue polyalanine helix in the EMR with dihedral angle values of $\phi = -57°$ and $\psi = -47°$. The hydrogen bonding energy calculated as a sum of electrostatic and steric van der Waals interactions was found to be -1.711 kcal/mol for all pairs of EMR units involved in a hydrogen bond. The distances between the participating carbonyl and amide EMR units were consistently 3.96 Å. In the retrieved full atomic representation, the carbonyl carbon and amide nitrogen atom distances were found to be 2.85 Å. The peptide dipoles are mostly aligned parallel to the helical axis, but rotate along the turn of the helix. These results are given in Table V, where it is shown that the distance between pairs of units involved in a hydrogen bond and the total energy of interaction remain unchanged, even though the dipoles rotate along the axis of the helix. The above results for the polyalanine α helix demonstrate that the energy functions accurately model directional hydrogen bonds.

We conclude that it is accurate to model the peptide unit as two dipole vectors, and hydrogen bonds as dipole–dipole interactions, which is possible in the EMR. The energy associated with directional interactions like hydrogen bonds can be calculated accurately using the energy functions of the EMR. The results in Table V indicate that for EMR units in a hydrogen bond, electrostatic effects are significantly larger than steric effects. Thus, the formation of regular structures such as a right-handed α helix depends strongly on the presence of favorable electrostatic interactions, and much less so on packing considerations.

EMR AND SIMULATING CHAIN DYNAMICS

A better understanding of the mechanisms involved in protein folding, structure formation, and functionality of proteins requires a detailed study of the large-scale dynamics in proteins. Internal motions in globular proteins have been well documented.[1] These vary from small-scale atomic fluctuations to larger scale rigid body motions of secondary structure elements. Local motions such as atomic fluctuations and side-chain and loop motions contribute to substrate binding and provide pathways for the entry and exit of ligands. Larger scale motions such as secondary structure transitions, hinge bending of domains, and coupled structural changes are important to understand activation and transition processes that contribute

TABLE V.
Hydrogen Bond Energies for Interacting EMR NH and CO Pairs in a 10-Element Polyalanine Helix, Including Cartesian Coordinates and Dipole Moments of Peptide Units.

Residue Pair	x	y	z	EMR Distance (Å)	μ_x	μ_y	μ_z	Steric Energy (kcal / mol)
4 (NH)	4.08	2.93	3.10	3.96	−0.148	−0.10	−0.159	−0.065
1 (CO)	2.18	1.36	0.00		−0.443	−0.153	−0.485	
5 (NH)	5.05	5.49	2.44	3.96	−0.171	−0.128	−0.108	−0.065
2 (CO)	2.05	4.35	0.11		−0.513	−0.393	−0.200	
6 (NH)	4.06	7.03	4.59	3.96	−0.123	−0.166	−0.120	−0.065
3 (CO)	1.28	4.70	2.98		−0.248	−0.528	−0.343	
7 (NH)	5.25	5.72	6.79	3.96	−0.116	−0.125	−0.167	−0.065
4 (CO)	3.52	3.19	4.27		−0.295	−0.272	−0.545	

The ϕ and ψ angles are $-57°$ and $-47°$, respectively. It is shown that the orientation of the dipoles vary for each residue, while the interaction energies, the magnitude of the peptide unit moments, and the distance between the interacting units remain unchanged.

directly to protein folding pathways. These motions occur on time scales of 10^{-10} s to a few seconds.[1] Conventional molecular dynamics simulations using full atomic representations are generally restricted to picosecond time scales due to computational limitations, and thus cannot address many of the large-scale dynamical processes. In the EMR, the number of interaction centers is significantly reduced from the full atomic representation. A chain n residues long will have $4n$ EMR interaction centers, irrespective of the sequence. Unlike a full atomic representation, calculation times are unaffected by the types of residues in the chain. As the size of the chain grows, the advantages of the EMR become more important. Gronbech-Jensen and Doniach,[9] McCammon et al.,[10] Daggett et al.,[23] and Schneller and Weaver[32] used different simulation techniques to study helix-coil transition rates of synthetic polyalanine and polyvaline chains. Depending on the choice of representations, the total simulation time extended from a few hundred picoseconds to a few microseconds.

We performed preliminary dynamics simulations of the helix-coil transition behavior of three synthetic 10-residue polyglycine, polyalanine, and polyvaline chains. Glycine is the smallest side chain and is very flexible in a polypeptide chain. Valine, the largest side chain of the three, is not as flexible. We used a time step of 0.3 ps and obtained trajectories for a total simulation time of 0.6 ns (600 ps). The simulations were performed on a cluster of DEC alpha machines running the DEC OSF/1[41] operating system and a 90-MHz pentium PC[42] running the shareware operating system Linux.[43] Each 0.6-ns simulation took a total time of 4 min on the DEC alpha cluster and 5 min on the pentium PC. Phase diagrams of the time histories of all the ϕ and ψ angles for the three chains studied are plotted in Figure 7a-c. This figure is representative of the behavior of the backbone dihedral angles within the time frame of simulation. The ϕ and ψ angles at the start of the simulation were $\phi = -57°$ and $\psi = -47°$, respectively, for all three chains. The ϕ-ψ angles of all the residues unravel from the helical range during the first 50-100 ps of the simulation. The limited mobility of ϕ and ψ for residue 9, measured by the time variation of these angles, is apparent in the figures. In all three chains that we studied, residue 9, which is closest to the C-terminus of the helix, remains in the helical range for most of the 600-ps simulation time. The time histories of the dihedral angles show that residue 9 is the first to return to the helical range, followed by residue 8, residue 7, and succeeding residues in the sequence. The time histories in Figure 7a-c clearly shows that helix formation starts with a residue at one end of the chain (residue 9 near the C-terminus for all three chains) and is followed by successive residues in the sequence. All 10 residues in each of the chains return to the helical range within a time span of 300 ps. This time span is similar for all three chains.

During the simulation, conformational flexibility of a chain is determined by the side-chain type, larger side chains like valine being less flexible compared to glycine and alanine. This is demonstrated by behavior of the ϕ and ψ angles shown in the time evolved phase plot of Figure 7a-c. Comparing the results of Figure 7a-c with the steric plots in Figures 4-6 shows that the trajectory of a residue in ϕ-ψ space is confined to allowed regions of the corresponding Ramachandran plot. The increased mobility of glycine side-chain residues is in accordance with the symmetric Ramachandran plot for this residue. Trajectories of alanine and valine side-chain residues are confined to well-defined helical and β-sheet secondary structure regions. A detailed analysis of much longer simulations of short segments of polypeptide chains will be presented elsewhere.

In addition we performed a 1.5-ns simulation of the helix-coil transition behavior of the 21 residue E helix obtained from the crystal structure of sperm whale myoglobin.[44] Myoglobin is an α-helical protein, the 3-dimensional structure of which is determined by the packing of the eight helices in the protein. Previous studies have indicated that the E helix does not interact substantially with the other seven helices.[45] This helix is therefore a good example of a naturally occurring isolated heteropolymeric polypeptide chain. The purpose of our investigation was to study the effect of the EMR parametrization on the dynamical properties of the E helix. Since the sequence includes a good cross section of the naturally occurring amino acids, a study of the dynamics is also a good test of the different polar and nonpolar EMR side chains. The sequence of the E helix used in the simulation is shown below. The numbers of the residues in the sequence are identical to the sequence of the E helix in the crystal structure of myoglobin.[44] E helix (21 residues): (58-78) [NT-S-D-D-L-K-K-H-G-V-T-V-L-T-A-L-G-A-I-L-K-K-CT]. Internal coordinates for the starting conformation of the helix were obtained from the crystal structure of sperm whale myoglobin.[44] Results of the 1.5-ns simula-

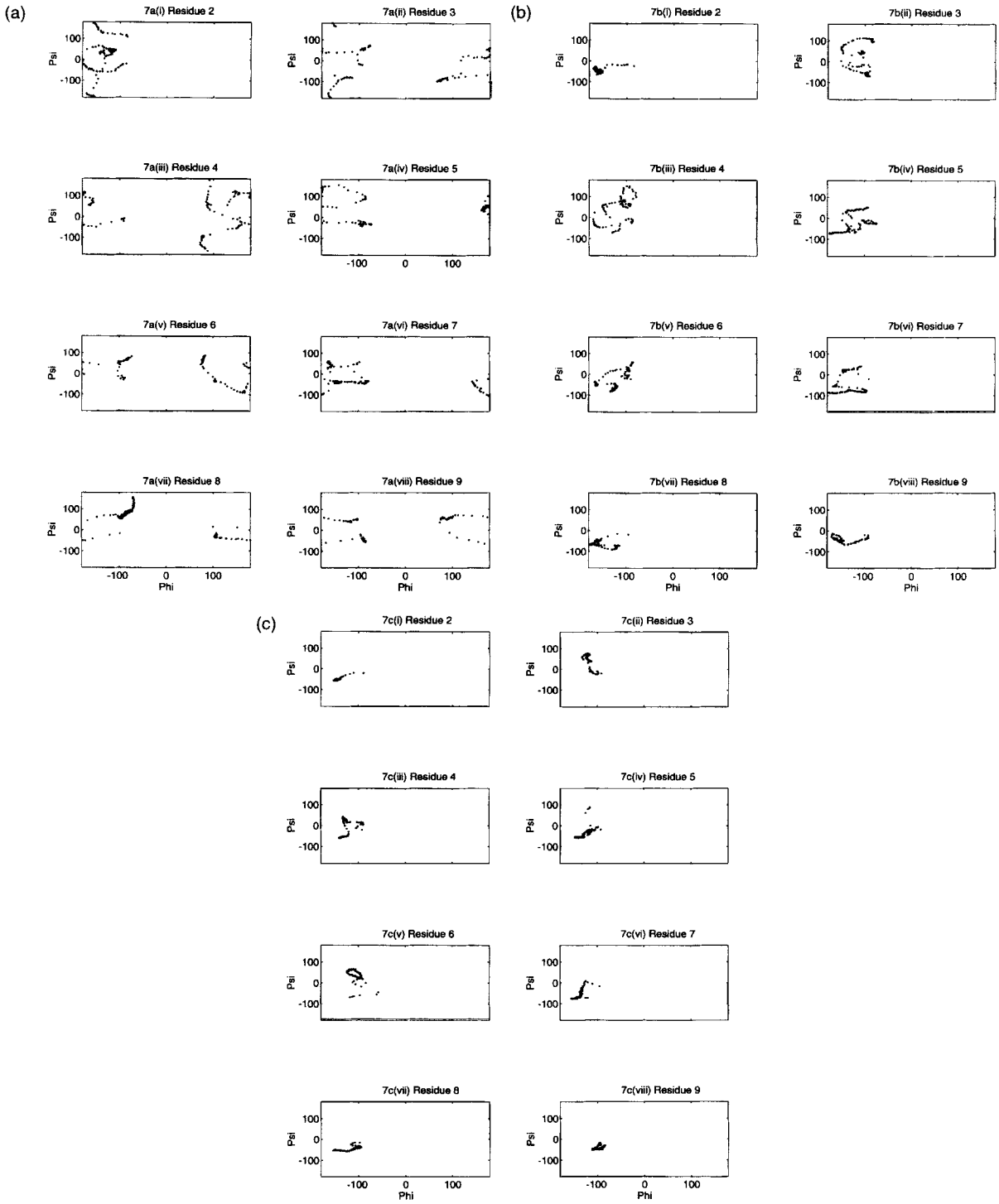


FIGURE 7. 600-ps phase-diagram time histories of the backbone ϕ, ψ dihedral angle pairs for 10-residue synthetic polyglycine, polyalanine, and polyvaline helical chains, respectively. The residue number is indicated for each phase plot. It must be noted that in this simulation all EMR units were allowed to move. Valine is not very flexible, and alanine possesses intermediate flexibility, while glycine is the most flexible. This is directly related to the sizes of the three side chains. The time histories for each (ϕ, ψ) pair of angles indicate that they move in and out of the helical region depending on the size of the amino acid residue in the chain and the location of the residue in the chain (i.e., near the N-terminus or near the C-terminus).

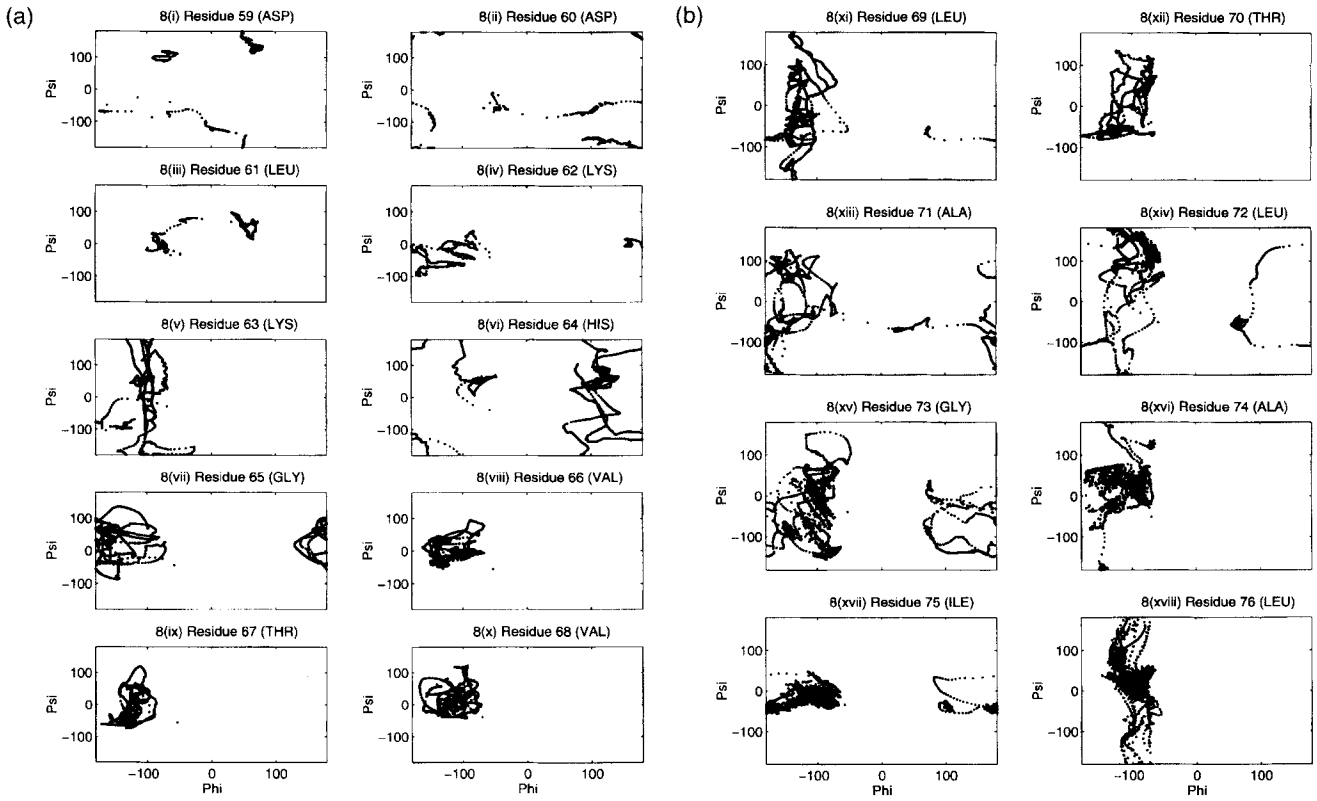


FIGURE 8. 1.5-ns phase space time histories of the backbone ϕ, ψ dihedral angle pairs for a 21-residue sequence of the E helix of sperm whale myoglobin. The residue number and type is indicated for each of the phase plots. Residues near the N-terminus have greater mobility compared to the other residues in the helix. The starting conformation was obtained from the crystal structure of sperm whale myoglobin.[43] The phase space trajectories for all the residues sample only the sterically allowed regions of the $\phi-\psi$ conformational space. Residues near the middle of the sequence prefer the helical conformation. This is indicative of partial helix formation. The charged residues (Asp and Lys) near the N-terminus are considerably more flexible. Lysine prefers a fully extended conformation as is evident from the phase space trajectories.

tion obtained using a time step of 0.01 ps are shown in Figures 8 and 9.

The sequence of the E helix indicates the absence of a well-defined hydrophobic core, i.e., the absence of a sequence of four or more adjacent hydrophobic residues in the sequence. The time evolved dynamical behavior of each of the residues is shown in detailed $\phi-\psi$ phase space diagrams in Figure 8. The dynamical trajectories indicate that in isolation the myoglobin E helix is partially formed. Residues near the N-terminus (residues 58–65) of this long chain tend to be more flexible and sample a wider region of conformational space. These residues (Asp 59, 60, Lys 62, 63, and His 64) are either highly charged or significantly polar groups, see Table II. The effect of the electrostatic interactions is bound to make them significantly more mobile than other nonpolar (hydrophobic) groups. Residues in the middle and near the C-terminus exhibit more systematic excursions in and out of the helical conformation. Hydrophobic residues such as valine (66, 68), alanine (71), leucine (69, 72, 76), and isoleucine (75) tend to prefer the helical conformation in this sufficiently long simulation. Threonine (67, 70), a hydroxyl group, also prefers the helical conformation. This could be due to the helix stabilizing interactions (resulting from the EMR force field) between its hydrophobic neighbors in the sequence.

Figure 9a is a plot of the angular variance of the $\phi-\psi$ angles of the different residues in the sequence. Again, residues near the end of the helix tend to be more flexible than the residues in the middle of the sequence, as is evident from the variation of the $\phi-\psi$ angles. Residues in the middle of the sequence are generally confined to the

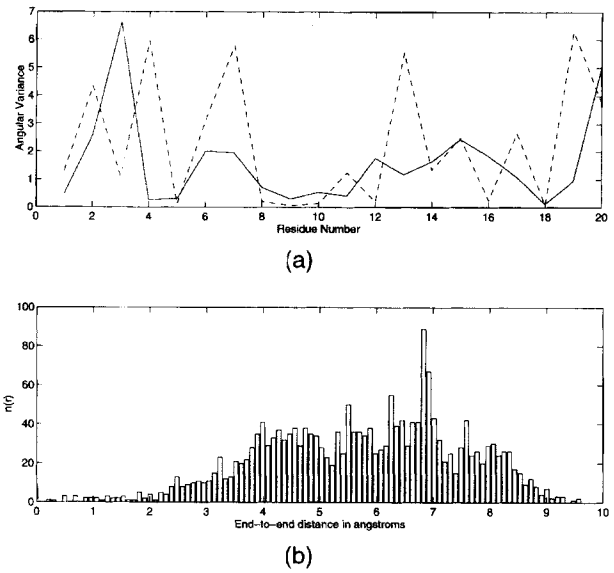


FIGURE 9. (a) Angular variance of the $\phi-\psi$ angle pairs for residues of the myoglobin E helix. The solid curves are the ϕ angles and the dashed curves are the ψ angles. Variations of the $\phi-\psi$ angles show close correlations in the E helix. (b) Distribution of the end to end distance of the E helix during the 1.5-ns simulation.

helical range, while residues near the terminal groups are considerably more flexible, indicating partial helix formation during the 1.5-ns simulation time. The end to end distance variation over the length of the simulation is also indicative of partial helix formation (see Fig. 8b). Detailed discussions of longer simulations of different heteropolymeric sequences using the EMR algorithm will be published elsewhere.

In general, the results of the E-helix simulation indicate that:

1. The EMR parameters and force field reflect the diverse nature of the different amino acids in a polypeptide sequence.
2. Isolated segments of EMR heteropolymeric chains are capable of forming partial secondary structures.
3. The structural and dynamical behavior of EMR polypeptide sequences is determined by the nature of the residues in the sequence and the intrachain interactions between these amino acids.
4. The EMR algorithm is useful for obtaining significant information about the early folding dynamics of small peptides through extended simulations of dynamical properties.

ACCURACY OF GEOMETRY AND ENERGY FUNCTIONS

We developed a representation of the polypeptide chain based on an electrostatic multipole description. In methods that truncate the complexity of a polypeptide chain, it is necessary that the description should accurately reflect the geometry of the chain. In addition, it is essential that we effectively simulate the stabilizing interactions that contribute to structure formation and long-range order. Inadequate potentials, despite geometric accuracy, result in an inaccurate description of the polypeptide chain. The three main requirements we sought to satisfy in developing the multipole representation were as follows.

1. The geometry of the EMR polypeptide backbone should be accurate, i.e., should be near-native. The ϕ and ψ angles calculated in the EMR description are in fact the native backbone dihedral angles in an all-atom representation.
2. The energy functions adequately model all the pairwise interactions in the molecule. The conformational energy maps, and hydrogen bond calculations validate the accuracy of the energy functions and the parameters used in these functions.
3. To be complete, it is essential that we be able to retrieve accurate full atomic informations from the EMR description. All the internal coordinates in the full atomic representation are preserved and retrieved from the EMR description with minimal deviations in peptide geometry.

These requirements were satisfied through a rational choice of units for the EMR description of the polypeptide, the use of simple transformations between coordinate representations, and accurate characterization of the hierarchy of important electrostatic interactions.

Future Directions

The diffusion collision model of protein folding[4,5] views the protein as composed of several parts (elementary microdomains), each short enough for all conformational alternatives to be searched through rapidly (compared to the folding time scale). The shortness of the segments implies that they are fluctuating among many conforma-

tions. The formation of tertiary structure occurs by the coalescence of these diffusing segments. The detailed dynamics and questions regarding the folding pathways depend on the properties of the microdomains. In this context, recent work has been done on simulating the secondary structure–coil transition properties of β sheets[46] and α helices[32] using the simplified whole residue approach proposed by McCammon et al.[10] Work has also been done to calculate the time of coalescence for two microdomains according to the diffusion collision model using both the simplified whole residue approach[47] and by modeling the microdomains as a pair of soft interacting spheres connected by an intervening string.[48]

The multi-microdomain coalesced aggregate in the DC model is a loosely associated molten-globule-like state of the protein.[5] Formation of such a molten globule is possible either via a diffusive mechanism or a hydrophobic (heteropolymeric) collapse mechanism suggested by Dill and coworkers.[49] A detailed investigation of the mechanism of molten-globule formation is necessary in order to the test the postulates of phenomenological folding models, either by Brownian dynamics simulations or Monte Carlo methods.[50] In the EMR a chain n residues long will have $4n$ interaction centers irrespective of side-chain types. The EMR algorithm allows for the generation of the full atomic representation, with fixed side-chain orientations, at the molten-globule stage after simulating the early folding dynamics with the EMR.

Recently Fezoui et al.[51] synthesized a *de novo* α-helical hairpin peptide containing two 17-residue helices connected by a 4-residue turn. The design of the helical hairpin peptide $\alpha t\alpha$,[51] provides a model system to investigate the nature of forces stabilizing secondary structures and the kinetics of helix association. The $\alpha t\alpha$ helical hairpin motif contains two elementary microdomains, i.e., the individual helices, and may be used as a relatively simple experimental system to study aspects of the DC and collapse models for protein folding. The role of hydrophobic residues in localizing secondary structure and the characterization of specific interactions is being studied experimentally through systematic point mutations of the original *de novo* designed $\alpha t\alpha$ peptide. A compact molten globule with partially formed secondary structure appears to form in a wide range of proteins within a millisecond time scale.[52,53] Since these intermediates form rapidly, within the dead time of stopped-flow studies, it has been difficult to investigate their folding kinetics. Constraints of computational time have prevented an investigation of their kinetics via simulation methods. In parallel with the experimental studies, kinetics of the association of helices can be obtained through EMR dynamical simulations, and Monte Carlo simulations as proposed by Sali and co-workers,[50] with modifications to include the EMR algorithm and force field. The EMR will be used for such an $\alpha t\alpha$ motif with smaller helices to study the effect of interhelical interactions including helix–coil transitions for the individual helices. Results of preliminary dynamics of short helices indicate that it is possible to obtain time histories for polypeptides containing different side chains in reasonable simulation times. The simulations will be performed for extended times to obtain details of the kinetics of helix formation and helix–helix interactions in an $\alpha t\alpha$ peptide. Results of such a simulation can be compared with the postulates of existing phenomenological models for protein folding.

References

1. C. L. Brooks III, M. Karplus, and B. M. Pettitt, *Adv. Chem. Phys.*, **71** (1988).
2. C. B. Anfinsen, *Science*, **181**, 223 (1973).
3. F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, *J. Mol. Biol.*, **112**, 535 (1977).
4. M. Karplus and D. L. Weaver, *Nature*, **260**, 401 (1976).
5. M. Karplus and D. L. Weaver, *Protein Sci.*, **3**, 650 (1994).
6. S. K. Burley and G. A. Petsko, *Adv. Protein Chem.*, **39**, 125 (1988).
7. T. Head–Gordon and C. L. Brooks III, *Biopolymers*, **31**, 77 (1990).
8. M. L. Levitt, *J. Mol. Biol.*, **101**, 57 (1976).
9. N. Gronbech–Jensen and S. Doniach, *J. Comput. Chem.*, **15**, 997 (1994).
10. J. A. McCammon, S. H. Northrup, M. Karplus, and R. M. Levy, *Biopolymers*, **19**, 2033 (1980).
11. P. Flory, Statistical Mechanics of Chain Molecules, Wiley, New York, 1969.
12. T. E. Creighton, Proteins: Structure and Molecular Properties, W. H. Freeman and Company, New York, 1983.
13. A. E. Wada, *Adv. Biophys.*, **9**, 1 (1976).
14. W. G. J. Hol, P. T. van Duijen, and H. J. C. Berendsen, *Nature*, **273**, 443 (1978).
15. R. P. Sheridan and L. C. Allen, *Biophys. Chem.*, **11**, 133 (1980).
16. N. E. Zhou, C. M. Kay, and R. S. Hodges, *J. Mol. Biol.*, **237**, 500 (1994).
17. G. N. Ramachandran and V. Sasisekharan, *Adv. Protein Chem.*, **23**, 283 (1993).
18. B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, *J. Comput. Chem.*, **4**, 187 (1983).

19. H. Goldstein, Classical Mechanics, Addison-Wesley, Reading, MA, 1980, chap. 4.
20. J. D. Jackson, Classical Electrodynamics, Wiley, New York, 1975, chaps. 3, 4.
21. W. L. Jorgensen, *J. Am. Chem. Soc.*, **77**, 4156 (1982).
22. W. L. Jorgensen and J. Tirado-Rives, *J. Am. Chem. Soc.*, **110**, 1657 (1988).
23. V. Daggett, P. Kollman and I. Kuntz, *Biopolymers*, **31**, 285 (1991).
24. B. E. Hingerty, R. H. Ritchie, T. L. Ferrell, and J. E. Turner, *Biopolymers*, **24**, 427 (1985).
25. J. C. Slater and J. G. Kirkwood, *Phys. Rev.*, **37**, 682 (1931).
26. K. S. Pitzer, *Adv. Chem. Phys.*, **2**, 59 (1959).
27. R. A. Scott and H. A. Scheraga, *J. Chem. Phys.*, **42**, 2209 (1965).
28. J. O. Hirschfelder, C. F. Curtis, and B. R. Bird, Molecular Theory of Gases and Liquids, Wiley, New York, 1954, chap. 13.
29. D. R. Lide, Ed., Handbook of Chemistry and Physics, 72nd ed., CRC Press, Boca Raton, FL, 1991-1992.
30. L. Pauling, Nature of the Chemical Bond, Cornell University Press, Ithaca, NY, 1960, chap. 7.
31. J. A. McCammon and S. C. Harvey, Dynamics of Proteins and Nucleic Acids, Cambridge University Press, New York, 1987.
32. W. Schneller and D. L. Weaver, *Biopolymers*, **33**, 1519 (1993).
33. W. F. van Gunsteren and H. J. C. Berendsen, *Mol. Phys.*, **34**, 1311 (1977).
34. D. L. Ermak and J. A. McCammon, *J. Chem. Phys.*, **69**, 1352 (1978).
35. W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, Numerical Recipes in C: The Art of Scientific Computing, Cambridge University Press, New York, 1992.
36. H. A. Scheraga, *Adv. Phys. Org. Chem.*, **6**, 103 (1968).
37. S. S. Zimmerman, M. S. Pottle, G. Nemèthey, and H. A. Scheraga, *Macromolecules*, **10**, 1 (1977).
38. L. Pauling, R. Corey, and H. R. Branson, *Proc. Natl. Acad. Sci. USA*, **37**, 205 (1951).
39. J. E. Alter, R. H. Andreatta, G. T. Taylor, and H. A. Scheraga, *Macromolecules*, **6**, 564 (1973).
40. J. M. Scholtz, S. Marqusee, R. L. Baldwin, E. J. York, J. M. Stewart, M. Santoro, and D. W. Bolen, *Proc. Natl. Acad. Sci. USA*, **88**, 2854 (1991).
41. Alpha farm consisting of four DEC 3000/800 servers. Runs the 1995 release of DEC OSF/1 3.0. CPU is a DEC 21064 200-MHz chip per each node with a total of four nodes.
42. Intel Pentium 90-MHz, 32 MRAM, 512 K cache.
43. L. Torvalds, Linux version 1.2.0, 1995.
44. S. E. Phillips, *Nature*, **273**, 247 (1978).
45. D. L. Weaver, *Biopolymers*, **32**, 477 (1992).
46. K. Yapa, D. L. Weaver, and M. Karplus, *Proteins: Struct., Function Genet.*, **12**, 237 (1992).
47. S. Lee, M. Karplus, D. Bashford, and D. Weaver, *Biopolymers*, **26**, 481 (1987).
48. D. L. Weaver, *Biopolymers*, **21**, 1275 (1982).
49. K. A. Dill, S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas, and H. S. Chan, *Protein Sci.*, **4**, 561 (1995).
50. A. Sali, E. Shakhnovich, and M. Karplus, *J. Mol. Biol.*, **235**, 1614 (1994).
51. Y. Fezoui, D. L. Weaver, and J. J. Osterhout, *Proc. Natl. Acad. Sci. USA*, **91**, 3675 (1994).
52. K. Kuwajima, *Proteins: Struct., Function Genet.*, **6**, 87 (1989).
53. O. B. Ptitsyn, R. H. Pain, G. V. Semisotnov, E. Zerovnik, and O. I. Razgulayev, *FEBS Lett.*, **164**, 21 (1990).